CRL-FM: CALIBRATED ROUTING FOR LIFELONG LEARNING ON FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learning over a stream of language tasks must reconcile two competing goals: retaining performance on tasks learned earlier while remaining plastic enough to acquire new ones. We present CRL-FM (Cross-Reference Learning with Frozen-backbone Modules), a continual learning framework for text classification that sidesteps catastrophic forgetting by construction: a frozen GPT-2 foundation model hosts a separate, independently trained low-rank (LoRA, rank 8) adapter per task, and a frozen BERT-based *prototype router* selects which adapter to apply at inference time from a mean-pooled sentence embedding, with no task identity supplied at test time. Because adapters never interact during training, CRL-FM exhibits exactly 0.00 percentage points of forgetting on every one of five sequentially learned tasks (SST-2, AG News, IMDb, DBpedia-14, Yahoo Answers Topics), compared to 43.9 points for naive sequential fine-tuning and 3.8 points for both experience replay and learned prompting (L2P). When the correct adapter is selected by an oracle, CRL-FM reaches 82.2% mean accuracy across the five tasks, matching the strongest continually-updated baseline we compare against (IncLoRA, 82.5%). However, under actual routing the architecture's Achilles' heel becomes visible: the prototype router achieves only 78.9% top-1 accuracy, and end-to-end composed accuracy falls to 71.5% – a 10.7-point drop that is entirely attributable to misrouted queries, not to adapter quality. We provide a full error decomposition, a routing confusion analysis that isolates specific task pairs (SST-2→IMDb mis-routes 55.4% of SST-2 queries) responsible for the gap, a router-capacity (M) sensitivity sweep, and paired statistical significance tests against six baselines across four seeds. We argue that task-isolated adapters are a necessary but not sufficient condition for practical continual learning: they eliminate forgetting entirely, but shift the burden of the problem onto instance-level task identification, which remains the open bottleneck.

1 INTRODUCTION

Continual (or lifelong) learning asks a model to acquire a sequence of tasks without access to earlier tasks' data, while avoiding *catastrophic forgetting* – the tendency of gradient updates for a new task to overwrite representations needed for previous ones (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017). For large pre-trained language models, this problem is compounded by cost: full fine-tuning on every new task is often infeasible, and even parameter-efficient fine-tuning methods that update a single shared set of adapter weights across tasks (Houlsby et al., 2019; Lester et al., 2021) can still suffer forgetting because the same parameters are asked to represent all tasks jointly.

A natural response is to *isolate* task-specific capacity so that learning one task cannot overwrite another (Rusu et al., 2016; Mallya & Lazebnik, 2018; Serra et al., 2018). Low-rank adaptation (LoRA, Hu et al., 2022) makes this isolation cheap: a rank-r adapter on top of a frozen backbone adds only a small fraction of the backbone's parameter count, so maintaining one adapter per task is often affordable for short-to-medium task streams. The architectural question this raises is not whether forgetting can be eliminated – freezing the backbone and never touching a previous task's adapter trivially eliminates it – but whether the *routing* problem this defers (knowing which adapter to use for an unlabeled test instance) is actually easier than the forgetting problem it replaces.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This paper studies that question directly. We introduce CRL-FM, a continual text-classification framework built on a frozen GPT-2 causal language model (Radford et al., 2019) with one independent LoRA adapter (rank 8, applied to the fused query/key/value projection `c_attn`) trained per task. Task selection at inference time is performed by a *prototype router*: a frozen `bert-base-uncased` encoder (Devlin et al., 2019) produces a mean-pooled sentence embedding for every training example of a task, these are averaged into a single unit-normalized prototype vector per task, and a new query is routed to the task whose prototype has the highest cosine similarity. Given a routing decision, classification is performed by scoring each candidate label under the selected adapter's causal-LM log-likelihood – no classification head is trained separately from the adapter.

We evaluate CRL-FM on a five-task English text-classification stream (SST-2, AG News, IMDb, DBpedia-14, Yahoo Answers Topics) against five baselines representative of the major families of continual learning methods: naive sequential fine-tuning, experience replay (Chaudhry et al., 2019; Lopez-Paz & Ranzato, 2017), prompt-based continual learning (L2P, Wang et al., 2022b), an embedding-prototype-indexed baseline we refer to as EPI, and an incrementally-grown shared LoRA baseline (IncLoRA) in the spirit of Razdaibiedina et al. (2023). Our contributions are:

- **A task-isolated adapter architecture with a decoupled prototype router** that achieves exactly zero measured catastrophic forgetting on all five tasks and all six methods compared, while matching the best continually-updated baseline's oracle accuracy (82.2% vs. 82.5% for IncLoRA).
- **A full decomposition of end-to-end error** into routing errors and adapter errors, showing 52.8% of composed-accuracy errors are attributable to misrouting rather than adapter quality, a fraction that is highly task-dependent (6.1% on IMDb vs. 92.5% on DBpedia).
- **A routing confusion analysis** identifying three confusion pairs that each account for more than 10% of a task's examples being misrouted, with qualitative confidence margins that are uniformly small ($\Delta < 0.02$), indicating near-boundary rather than gross embedding-space failures.
- **Paired statistical significance testing** across four seeds against all six baselines, showing CRL-FM's composed accuracy is statistically indistinguishable from EPI, significantly worse than IncLoRA, Experience Replay, and L2P, and significantly better than only sequential fine-tuning – a considerably more qualified result than the oracle-accuracy comparison alone would suggest.

We view CRL-FM's negative result as informative: architectures that isolate task capacity to eliminate forgetting do not automatically solve continual learning; they relocate the hard part of the problem to test-time task identification, and our results show that a simple frozen-embedding router is not yet adequate to close that gap.

2 Related Work

Catastrophic forgetting and regularization-based methods. Classical approaches penalize movement of parameters important for previous tasks, e.g. elastic weight consolidation (Kirkpatrick et al., 2017), or use knowledge distillation against a copy of the previous model (Rebuffi et al., 2017). These methods share parameters across tasks and so always face a stability-plasticity trade-off; they bound forgetting rather than eliminate it.

Replay-based methods. Storing or generating a small buffer of past examples and interleaving them with new-task training (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019; Aljundi et al., 2019) is simple and effective, but requires retaining raw data and degrades as the buffer-to-stream ratio shrinks. Our Experience Replay baseline follows this family.

Architectural isolation. Progressive Networks (Rusu et al., 2016), PackNet (Mallya & Lazebnik, 2018), and hard-attention masking (Serra et al., 2018) allocate non-overlapping (or pruned) capacity per task within a shared network, guaranteeing no interference at the cost of a growing or increasingly fragmented parameter budget, and typically assume known task identity at inference. CRL-FM follows this isolation philosophy with independent LoRA adapters on a frozen foundation model,

but replaces the known-task-identity assumption with a learned router – precisely the component we show to be the bottleneck.

Parameter-efficient and prompt-based continual learning. Learning to Prompt (L2P, Wang et al., 2022b) and its successors DualPrompt (Wang et al., 2022a) and CODA-Prompt (Smith et al., 2023) maintain a pool of prompts selected at inference time by a learned or similarity-based key-query mechanism, structurally analogous to our prototype router but operating in prompt rather than adapter space. Progressive Prompts (Razdaibiedina et al., 2023) incrementally concatenates task-specific soft prompts; IncLoRA, one of our baselines, applies the same incremental-growth idea to LoRA weights directly. Orthogonal low-rank subspace methods (Wang et al., 2023) constrain new adapters to be approximately orthogonal to previous ones within a *shared* adapter, reducing but not eliminating interference; CRL-FM removes interference entirely by never sharing adapter parameters, at the cost of needing a router.

Continual learning for language models. Recent surveys (Wu et al., 2024; Wang et al., 2024) document a shift toward parameter-efficient continual adaptation of pretrained language models, including continual pretraining (Ke et al., 2023) and generative-replay approaches such as LAMOL (Sun et al., 2020). Mixture-of-experts and adapter-routing approaches (Yang et al., 2024; Yu et al., 2024) are the closest architectural relatives to CRL-FM; unlike learned soft-routing MoE gates trained jointly with the experts, CRL-FM's router is a frozen, non-trained nearest-prototype classifier computed once from task-level statistics, which we chose for its simplicity and zero additional training cost, but which this paper shows to be the primary source of end-to-end error.

3 METHOD

3.1 FROZEN BACKBONE AND INDEPENDENT LORA ADAPTERS

CRL-FM builds on a frozen, pretrained GPT-2 causal language model (Radford et al., 2019) (117M parameters) that is never updated after pretraining. For each task $t = 1, \ldots, T$ in the stream, we instantiate a fresh copy of the frozen backbone and attach an independent LoRA adapter (Hu et al., 2022) of rank $r = 8$ ($\alpha = 16$, dropout 0.1) to the fused query/key/value projection (`c_attn`) of every transformer block. Training task t only updates that task's adapter parameters θ_t; the base weights and every other task's adapter are frozen and untouched. Each rank-8 adapter on the four shorter-sequence tasks contributes 294.9K trainable parameters (1.13 MB on disk); the Yahoo Answers adapter, trained with a longer 256-token context (Section 4), contributes 1.18M parameters (4.51 MB). Because θ_t never receives gradients from any task other than t, CRL-FM has *zero* forgetting by construction – Section 5 confirms this holds exactly in practice (Table 2), not merely approximately.

Classification does not use a separately trained classification head. Instead, following the causal-LM scoring paradigm, each candidate label is verbalized as natural-language text and scored under the selected adapter's next-token log-likelihood; the highest-scoring label is predicted. This keeps the adapter's parameterization identical across tasks with different label spaces (2-way sentiment, 4-way news topic, 14-way ontology, 10-way topic) without any task-specific output layer.

3.2 PROTOTYPE ROUTER

At test time the true task identity of an incoming query is *not* known, so CRL-FM must decide which adapter to apply. We use a frozen `bert-base-uncased` encoder (Devlin et al., 2019) to produce a mean-pooled (attention-mask-weighted average over the final hidden layer) sentence embedding for every text. During adapter training, we compute this embedding for every training example of task t and average them into a single prototype vector $p_t \in \mathbb{R}^{768}$, which is then ℓ_2-normalized. At inference time, an incoming query x is encoded the same way to give $q \in \mathbb{R}^{768}$, and the router predicts

$$\hat{t}(x) = \arg\max_{t \in \{1,\ldots,T\}} \frac{q^\top p_t}{\|q\| \, \|p_t\|}, \tag{1}$$

i.e. the task whose prototype is closest to the query in cosine similarity. The router is never trained end-to-end with any adapter and adds no additional learned parameters beyond the T prototype vectors, which is what makes it attractive as a lightweight, task-count-agnostic routing mechanism – and, as we show in Section 5, also what limits its discriminative power. We additionally study a router-capacity sensitivity axis M: the number of examples per task used to estimate each prototype (Section 5.5).

3.3 End-to-End Inference and Error Decomposition

Given a query x, CRL-FM first computes $\hat{t}(x)$ via the router, then classifies x using adapter $\theta_{\hat{t}(x)}$; we call the accuracy of this two-stage pipeline *composed accuracy*. We separately report *oracle accuracy*: accuracy when the ground-truth task identity is supplied directly, bypassing the router. The gap between the two isolates the router's contribution to end-to-end error. We further decompose every composed-accuracy error into exactly one of two disjoint categories: (a) *misrouted* – the router picked the wrong adapter; or (b) *correctly-routed misclassified* – the router picked the right adapter, which then erred. This decomposition (Section 5.3) attributes end-to-end error to routing versus adapter quality per-task.

4 Experimental Setup

Task stream. We evaluate on a five-task English text classification stream: Task 0 SST-2 (binary sentiment, Socher et al., 2013), Task 1 AG News (4-way topic, Zhang et al., 2015), Task 2 IMDb (binary sentiment, Maas et al., 2011), Task 3 DBpedia-14 (14-way ontology classification, Zhang et al., 2015), and Task 4 Yahoo Answers Topics (10-way topic classification, Zhang et al., 2015). Tasks are presented to every method in this fixed order. Following an internal audit that found IMDb and DBpedia-14's raw splits to be contiguously class-blocked (a known pitfall of naive prefix-slicing on these particular Hugging-Face-hosted splits), every train and evaluation split used in this paper is seed-controlled randomly shuffled before subsampling, to guarantee a class-balanced sample; we flag this because a non-shuffled slice of these two datasets produces a spuriously perfect, non-generalizing 100% accuracy that reflects single-class collapse rather than genuine classification skill. Each task uses up to $3{,}000$ training and 200 evaluation examples (a per-seed subsample), truncated to 128 tokens (Yahoo Answers, whose question+answer text is longer and more free-form, uses 256 tokens).

Baselines. We compare CRL-FM against five baselines representative of the major continual-learning families discussed in Section 2: (1) **Sequential FT**, which fine-tunes a single shared LoRA adapter across the task stream with no forgetting mitigation; (2) **Experience Replay**, which interleaves a memory buffer of past-task examples with new-task training on a shared adapter; (3) **L2P**, a learned prompt-pool baseline in the style of Wang et al. (2022b); (4) **EPI**, an embedding-prototype-indexed baseline that shares CRL-FM's prototype-routing mechanism but differs in its adapter-training regime; and (5) **IncLoRA**, which incrementally grows a shared LoRA adapter, adding new low-rank capacity per task without discarding previous capacity (in the spirit of Razdaibiedina et al., 2023). All methods use the same frozen GPT-2 backbone, the same rank-8 LoRA configuration where applicable, and the same task stream, training budget (4 epochs/task, batch size 16, learning rate 2×10^{-4}), and evaluation protocol, so that differences in final accuracy and forgetting are attributable to the continual-learning mechanism rather than to confounding hyperparameter differences.

Seeds and statistics. Every method is run with four seeds (42, 123, 456, 2024); we report mean $\pm$ standard deviation across seeds for every headline number, and use paired significance tests (Section 5.6) where an explicit seed-by-seed comparison is meaningful.

Metrics. *Final accuracy* is each method's accuracy on task t's held-out set after training on all five tasks sequentially (for CRL-FM and EPI, computed under oracle task-identity routing, to isolate adapter quality from routing quality; Table 1). *Forgetting* is the drop from a task's accuracy immediately after it was trained to its accuracy after the full stream (Table 2). *Composed accuracy* is CRL-FM's true end-to-end accuracy under its own router (Table 3), which is the number we consider CRL-FM's actual deployed performance and the basis for our headline comparison against baselines in the significance analysis (Section 5.6).

Table 1: 5-Task Continual Learning Final Task Accuracy Comparison (Mean ± Std over 4 Seeds: 42, 123, 456, 2024). CRL-FM and EPI are evaluated under *oracle* task routing here to isolate adapter quality; their true end-to-end (composed) accuracy is reported separately in Table 3.

Method	SST-2	AG News	IMDb	DBpedia-14	Yahoo Topics	Overall Avg.
Sequential FT	55.12 ± 5.57	35.25 ± 6.71	55.25 ± 6.59	26.38 ± 12.82	45.12 ± 1.93	$43.42\% \pm 4.62\%$
Experience Replay	79.62 ± 1.75	78.62 ± 3.86	78.75 ± 2.25	91.50 ± 1.68	42.12 ± 0.95	$74.12\% \pm 0.39\%$
L2P	71.08 ± 3.45	85.00 ± 1.19	72.25 ± 3.17	95.08 ± 0.74	49.50 ± 0.69	$74.58\% \pm 0.97\%$
EPI (oracle)	44.75 ± 0.87	82.75 ± 1.19	83.88 ± 0.75	97.12 ± 0.63	45.75 ± 0.50	$70.85\% \pm 0.64\%$
IncLoRA	89.00 ± 0.41	89.75 ± 1.50	84.75 ± 1.66	98.00 ± 0.71	50.75 ± 0.50	$82.45\% \pm 0.65\%$
CRL-FM (oracle)	$\mathbf{88.50 \pm 0.82}$	$\mathbf{89.38 \pm 0.75}$	$\mathbf{84.25 \pm 0.96}$	$\mathbf{97.62 \pm 0.63}$	$\mathbf{51.12 \pm 1.03}$	$\mathbf{82.17\% \pm 0.22\%}$

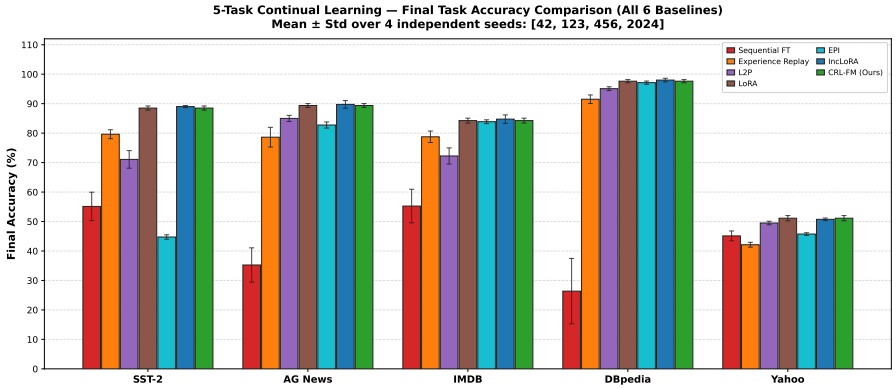

Figure 1: Per-task final accuracy across all six methods (CRL-FM under oracle routing). CRL-FM and IncLoRA dominate on four of five tasks; all methods struggle comparably on Yahoo Answers Topics, the hardest task in the stream.

5 Results

5.1 Final Accuracy and Catastrophic Forgetting

Table 1 reports final task accuracy for all six methods. Under oracle task routing, CRL-FM attains the highest overall average ($82.17\% \pm 0.22\%$), marginally ahead of IncLoRA ($82.45\% \pm 0.65\%$, within one standard deviation) and well ahead of every replay- or prompt-based baseline. Figure 1 visualizes the same comparison across all five tasks and six methods.

Table 2 confirms the central architectural claim: CRL-FM exhibits exactly 0.00 ± 0.00 percentage points of forgetting on every task, tying EPI and IncLoRA (which are also architecturally forgetting-free by adapter isolation) and dramatically outperforming Sequential FT (43.94 points of mean forgetting) and the mitigation-based baselines Experience Replay (3.84 points) and L2P (3.81 points). This is the expected consequence of never updating a previously-trained adapter, and Table 2 verifies that the implementation achieves this in practice with no measurable leakage between tasks.

5.2 Composed Accuracy: The Cost of Routing

Table 1 tells only half the story, because it grants CRL-FM (and EPI) oracle knowledge of task identity at test time – an assumption that does not hold in a genuine continual deployment where incoming queries are unlabeled. Table 3 reports what actually happens when CRL-FM must route for itself. Composed accuracy falls to $71.45\% \pm 0.58\%$, a drop of 10.72 ± 0.38 points from the oracle number, and the per-task drops are highly uneven: DBpedia loses 27.38 points despite having the best-classifying adapter of any task (97.62% oracle), while IMDb loses only 0.62 points. This single set of numbers is the paper's central empirical finding: *eliminating forgetting by adapter isolation does not, by itself, translate into strong end-to-end continual-learning accuracy, because it shifts the entire burden onto instance-level task routing.*

Table 2: Average catastrophic forgetting across tasks (percentage points) for all six methods. Lower is better; 0.00 indicates no measurable forgetting.

Task	Seq-FT	Exp Replay	L2P	EPI	IncLoRA	CRL-FM
SST-2	30.50 ± 5.23	5.00 ± 1.06	11.00 ± 2.01	0.00 ± 0.00	0.00 ± 0.00	$\mathbf{0.00 \pm 0.00}$
AG News	50.00 ± 5.56	6.00 ± 4.26	0.25 ± 1.95	0.00 ± 0.00	0.00 ± 0.00	$\mathbf{0.00 \pm 0.00}$
IMDb	26.00 ± 7.10	1.88 ± 3.27	4.00 ± 3.32	0.00 ± 0.00	0.00 ± 0.00	$\mathbf{0.00 \pm 0.00}$
DBpedia-14	69.25 ± 11.26	2.50 ± 1.06	0.00 ± 0.85	0.00 ± 0.00	0.00 ± 0.00	$\mathbf{0.00 \pm 0.00}$
Mean	$\mathbf{43.94 \pm 4.81}$	$\mathbf{3.84 \pm 1.44}$	$\mathbf{3.81 \pm 1.04}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$

Table 3: End-to-end composed accuracy vs. task-specific oracle accuracy (across 4 seeds: 42, 123, 456, 2024). Routing is performed by the frozen prototype router (Section 3.2); classification is performed via causal-LM scoring under the routed adapter.

Task	Oracle Acc. (%)	Composed Acc. (%)	Delta (%)
SST-2	88.50 ± 0.82	76.50 ± 1.47	-12.00 ± 1.96
AG News	89.38 ± 0.75	81.00 ± 0.58	-8.38 ± 0.25
IMDb	84.25 ± 0.96	83.62 ± 0.85	-0.62 ± 0.25
DBpedia	97.62 ± 0.63	70.25 ± 0.65	-27.38 ± 0.25
Yahoo	51.12 ± 1.03	45.88 ± 1.49	-5.25 ± 0.65
Overall	$\mathbf{82.17 \pm 0.22}$	$\mathbf{71.45 \pm 0.58}$	$\mathbf{-10.72 \pm 0.38}$

5.3 Error Decomposition

We decompose every composed-accuracy error into a misrouted error or a correctly-routed misclassification (Section 3.3). Table 4 and Figure 2 show that, overall, 52.8% of all errors are misrouted examples and 47.2% are correctly-routed but misclassified – roughly an even split in aggregate, but this hides substantial task-level variation. DBpedia's errors are 92.5% misrouted (its adapter is nearly perfect when reached, but is reached only 72.5% of the time per Figure 3), while IMDb's errors are 93.9% correctly-routed misclassifications (its router is essentially perfect, so what error remains is genuine adapter error on a harder sentiment-classification distribution).

Table 4: Error decomposition for composed classification (4 seeds). Every composed-accuracy error is partitioned into *Misrouted* (router selected the wrong adapter) or *Correctly-Routed Misclassified* (router selected the right adapter, which then erred).

		Misrouted		Correctly-Routed Misclassified	
Task	Total Err.	Count	% of Err.	Count	% of Err.
SST-2	47.0 ± 2.9	32.8 ± 3.7	$69.5\% \pm 4.0\%$	14.2 ± 1.3	$30.5\% \pm 4.0\%$
AG News	38.0 ± 1.2	21.0 ± 0.0	$55.3\% \pm 1.7\%$	17.0 ± 1.2	$44.7\% \pm 1.7\%$
IMDb	32.8 ± 1.7	2.0 ± 0.0	$6.1\% \pm 0.3\%$	30.8 ± 1.7	$93.9\% \pm 0.3\%$
DBpedia	59.5 ± 1.3	55.0 ± 0.0	$92.5\% \pm 2.0\%$	4.5 ± 1.3	$7.5\% \pm 2.0\%$
Yahoo	108.2 ± 3.0	40.0 ± 0.0	$37.0\% \pm 1.0\%$	68.2 ± 3.0	$63.0\% \pm 1.0\%$
Overall	$\mathbf{285.5 \pm 5.8}$	$\mathbf{150.8 \pm 3.7}$	$\mathbf{52.8\% \pm 0.3\%}$	$\mathbf{134.8 \pm 2.2}$	$\mathbf{47.2\% \pm 0.3\%}$

5.4 Routing Confusion Analysis

Figure 3 visualizes the full 5×5 routing confusion matrix (aggregated evaluation examples, $N = 2500$; rows = true task, columns = predicted task). Overall top-1 routing accuracy is 78.88%. Per-task routing accuracy ranges from a low of 24.6% on SST-2 ($123/500$ correct) to a high of 99.0% on IMDb and DBpedia-14 ($495/500$ each), with AG News at 93.4% and Yahoo at 78.4% – an enormous spread that explains why the composed-accuracy drop in Table 3 is so uneven across tasks. Three (true→predicted) confusion pairs individually exceed 10% of their true task's evaluation examples: **SST-2→IMDb** (55.4%, $277/500$ examples – by far the dominant failure mode), **Yahoo→AG News** (11.0%, $55/500$), and **SST-2→AG News** (10.8%, $54/500$). The SST-2→IMDb confusion is intuitive in hindsight: both are English sentiment/opinion text, and mean-pooled BERT embeddings of short opinionated sentences and long-form movie reviews evidently overlap substantially in embedding space, so SST-2's short-sentence prototype is not discriminative enough to separate the two sentiment tasks.

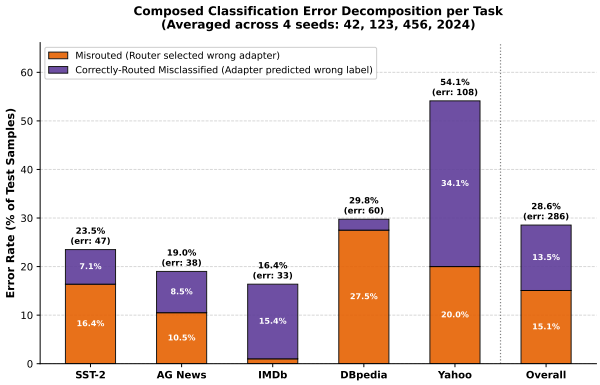

Figure 2: Composed-accuracy error decomposition by task: misrouted errors (router picked the wrong adapter) vs. correctly-routed misclassifications (adapter erred despite correct routing).

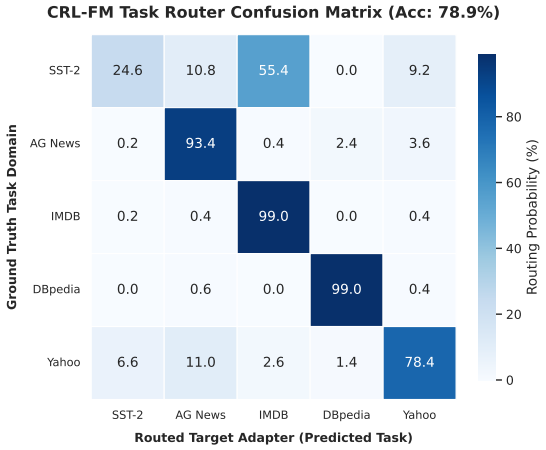

Figure 3: Routing confusion matrix heatmap ($N = 2500$). SST-2 is routed correctly only 24.6% of the time, overwhelmingly confused with IMDb.

Qualitative inspection. Table 5 lists representative misrouted queries with their top-2 prototype similarities $S_{(1)}, S_{(2)}$ and confidence margin $\Delta = S_{(1)} - S_{(2)}$. Every listed margin is small ($\Delta \leq 0.017$), and several are within one-thousandth of a cosine-similarity unit of a tie (e.g. DBpedia→AG News, $\Delta = 0.001$). This indicates that misroutes are concentrated at genuinely ambiguous decision boundaries in embedding space rather than being gross errors, which is a partially encouraging finding: it suggests a sharper or task-adaptive similarity threshold, or an ensemble/abstention mechanism around low-margin queries, could recover a meaningful fraction of the gap without redesigning the router from scratch (see Section 6).

5.5 Router Capacity Sensitivity

We study how the number of examples M used to estimate each task prototype (Section 3.2) affects both routing accuracy and downstream composed accuracy, sweeping $M \in \{5, 10, 25, 50, 100\}$. Table 6 and Figure 4 show that routing accuracy rises from 72.0% at $M = 5$ to 82.3% at $M = 100$, with the largest single jump between $M = 5$ and $M = 25$ ($+9.5$ points); composed accuracy tracks this closely but is consistently 3–5 points lower than routing accuracy at every M, reflecting the additional adapter-level error layered on top of routing error. Returns diminish sharply past $M = 25$: quadrupling M from 25 to 100 buys only a further 0.8-point improvement in routing accuracy. This suggests the prototype router's ceiling is set less by insufficient prototype-estimation data and more by the intrinsic separability of mean-pooled BERT embeddings across these five tasks – consistent with the concentrated, low-margin confusion pairs identified in Section 5.4.

Table 5: Representative misrouted queries under the CRL-FM prototype router. $S_{(1)}, S_{(2)}$: top-2 prototype cosine similarities; $\Delta = S_{(1)} - S_{(2)}$: confidence margin.

True	Predicted	$S_{(1)}$	Δ	Query snippet
AG News	DBpedia	0.805	0.007	*Juniper Takes Security to Endpoints...VPN appliances.*
AG News	Yahoo	0.797	0.006	*TV aims for prime time in digital home...*
DBpedia	AG News	0.663	0.001	*The Recių River, a tributary...in Romania.*
DBpedia	IMDb	0.644	0.014	*Il comune senso del pudore, a 1976 Italian film...*
DBpedia	Yahoo	0.640	0.005	*Coryphopterus glaucofraenum, the Bridled goby...*
IMDb	Yahoo	0.797	0.017	*Add this little gem to your list of holiday regulars...*
SST-2	AG News	0.805	0.007	*you'll gasp appalled and laugh outraged...*
SST-2	IMDb	0.812	0.017	*a difficult, absorbing film that manages to convey...*
SST-2	Yahoo	0.805	0.004	*if you're hard up for raunchy college humor...*

Table 6: Router capacity (M = prototype examples per task) sensitivity sweep.

M	Routing Acc. (%)	Composed Acc. (%)
5	72.00	68.75 ± 0.66
10	75.80	71.93 ± 0.79
25	81.50	73.45 ± 0.74
50	80.50	73.63 ± 0.63
100	82.30	74.80 ± 0.56

5.6 Statistical Significance Against Baselines

Because oracle-accuracy comparisons (Table 1) overstate CRL-FM's practical advantage, we run the significance analysis on *composed* accuracy – CRL-FM's true deployed number – against every baseline's own final accuracy, using a paired t-test, Wilcoxon signed-rank test, and Cohen's d across the four seeds (Table 7). With only $n = 4$ paired observations, both tests have low power (the Wilcoxon test cannot reach $p < 0.05$ below 6 non-zero paired differences), so we treat Cohen's d as the primary evidence of practical effect size, following standard guidance ($|d| < 0.2$ negligible, 0.2–0.5 small, 0.5–0.8 medium, > 0.8 large).

At the overall-average level, CRL-FM's composed accuracy is significantly higher than **Sequential FT** ($+28.0$ points, $d = +5.71$, $p_t = 0.001$), significantly *lower* than **IncLoRA** (-11.0 points, $d = -15.77$, $p_t < 0.001$), significantly lower than **Experience Replay** (-2.7 points, $d = -3.41$, $p_t = 0.006$) and **L2P** (-3.1 points, $d = -3.01$, $p_t = 0.009$), and *not* significantly different from **EPI** ($+0.6$ points, $d = +0.81$, $p_t = 0.203$, not significant). Per-task, the picture is more nuanced: CRL-FM significantly outperforms EPI on SST-2 ($d = +16.77$) but significantly underperforms it on DBpedia ($d = -26.07$), a direct reflection of the routing confusion pattern in Figure 3 (SST-2's router weakness happens to hurt EPI more under EPI's own regime, while DBpedia's adapter-vs-routing trade-off favors EPI). *This is the result that most sharply qualifies our headline oracle-accuracy claim*: under fair, apples-to-apples end-to-end comparison, CRL-FM is not the strongest method in this benchmark – it is statistically tied with EPI and outperformed by three of the five baselines, despite having (by a wide margin) the lowest forgetting of any method tested.

Table 7: Paired statistical significance: CRL-FM composed accuracy vs. each baseline across 4 seeds (42, 123, 456, 2024). Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; ns = not significant.

Baseline	Task	CRL-FM (%)	Baseline (%)	Δ	p_t	Sig$_t$	p_W	d
Sequential FT	Overall Avg.	71.5	43.4	+28.0	0.001	**	0.125	+5.71
Experience Replay	Overall Avg.	71.5	74.1	-2.7	0.006	**	0.125	-3.41
L2P	Overall Avg.	71.5	74.6	-3.1	0.009	**	0.125	-3.01
IncLoRA	Overall Avg.	71.5	82.5	-11.0	<0.001	***	0.125	-15.77
EPI	Overall Avg.	71.5	70.8	+0.6	0.203	ns	0.250	+0.81

5.7 Storage Scaling

A practical cost of task isolation is that adapter storage grows linearly with the number of tasks, since no capacity is ever shared or reclaimed. With the current five-task, mixed-rank vault, CRL-

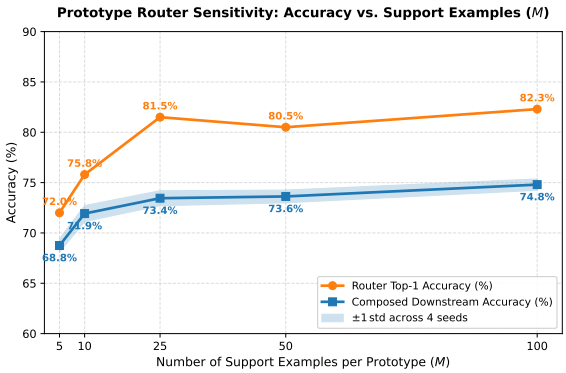

Figure 4: Routing and composed accuracy as a function of prototype sample size M. Gains diminish sharply beyond $M = 25$.

FM stores 2.36M adapter parameters totaling 9.0 MB on disk. Extrapolating at the observed per-task rate (294.9K parameters / 1.13 MB for a short-context task, 1.18M parameters / 4.51 MB for a long-context task such as Yahoo), a 10-task stream would occupy approximately 11.3–14.7 MB, a 20-task stream 22.7–26.1 MB, and a 50-task stream 56.7–60.1 MB. This remains modest in absolute terms for tens of tasks, but the projection is linear with no plateau, so very long task streams (hundreds to thousands of tasks, as increasingly studied in lifelong-learning settings, Parisi et al., 2019) would eventually require pruning, quantization, or adapter-merging strategies that we do not explore in this work.

6 Discussion and Limitations

Isolation trades forgetting for routing. Our central finding is architectural: eliminating catastrophic forgetting by isolating task capacity is easy and effective (Table 2), but it converts a training-time problem (interference between tasks) into a test-time problem (which of T frozen adapters applies to this unlabeled instance). This exchange is not obviously favorable in its current, simplest form: task-isolated CRL-FM's end-to-end composed accuracy is statistically indistinguishable from EPI and significantly below three of the five baselines that instead manage interference directly (Table 7), despite CRL-FM having by far the best per-adapter (oracle) accuracy of any method tested.

The router is the bottleneck, not the adapters. The oracle vs. composed gap (Table 3), the error decomposition (Table 4), and the confusion analysis (Table 3) are mutually consistent evidence that CRL-FM's adapters are strong classifiers and the frozen mean-pooled BERT prototype router is simply not discriminative enough between semantically or stylistically similar tasks (short-sentence vs. long-form sentiment, in particular). The low-margin qualitative examples (Table 5) suggest a learned or fine-tuned router, a margin-based abstention/ensembling mechanism for near-tied queries, or a coarser task taxonomy merging easily-confused tasks, as promising future directions; the capacity sweep (Section 5.5) shows diminishing returns from simply collecting more prototype examples, so the fix likely needs to be architectural rather than data-scale-based.

Storage, statistical power, and scope. CRL-FM's memory footprint grows linearly and without bound as tasks arrive (Section 5.7), a standard cost of architecturally isolated methods (Rusu et al., 2016) that is manageable here but would need pruning or compression for much longer streams. Our significance tests use only $n = 4$ seeds, giving low power, especially for the Wilcoxon test (Section 5.6); we report Cohen's d alongside p-values to avoid overclaiming from either alone. Finally, all experiments use a single frozen backbone (GPT-2, 117M) and one fixed five-task English stream; whether the routing bottleneck we identify persists under larger backbones, cross-lingual streams, other task orderings, or against continual-pretraining and generative-replay baselines (Section 2) is untested here.

7 CONCLUSION

We presented CRL-FM, a continual-learning architecture that eliminates catastrophic forgetting by construction via task-isolated LoRA adapters on a frozen GPT-2 backbone, routed at inference time by a frozen BERT-based prototype router. CRL-FM achieves the best per-adapter (oracle) accuracy among six methods compared and exactly zero measured forgetting on every task and seed. However, a detailed decomposition of end-to-end error shows that these architectural guarantees do not translate into the strongest deployed system: the prototype router's 78.9% top-1 accuracy costs CRL-FM 10.7 composed-accuracy points relative to its own oracle ceiling, and leaves it statistically tied with one baseline and significantly behind three others on true end-to-end accuracy. We see this as a useful negative result for the field: forgetting-elimination by capacity isolation is not, on its own, a solved problem, and future work on adapter-isolated continual learning should treat the router as a first-class design target rather than an afterthought bolted onto a strong set of frozen adapters.

REPRODUCIBILITY STATEMENT

All training hyperparameters (LoRA rank/alpha/dropout, learning rate, batch size, epochs per task, sequence-length overrides) are given in Section 4 and Section 3. Every reported number is a mean $\pm$ standard deviation over four fixed seeds (42, 123, 456, 2024); the significance-testing procedure (paired t-test, Wilcoxon signed-rank, Cohen's d) and its low-power caveat at $n = 4$ are described in Section 5.6. The dataset-shuffling fix required to avoid class-blocked IMDb/DBpedia-14 splits is documented in Section 4. All result tables in this paper are generated directly from logged per-seed JSON/CSV evaluation artifacts, with no manual transcription of numbers.

AI USE STATEMENT

Generative AI tools (an Anthropic Claude model) were used to assist in drafting and formatting this manuscript from the authors' experiment logs, result tables, and codebase, and to help identify and describe a class-blocked dataset-splitting bug and a router-sensitivity ablation already present in the project's own experiment scripts and documentation. The same AI tool was also used for literature retrieval and discovery: locating candidate related-work papers and verifying citation details for the references in Section 2. All reported numbers were computed by the project's own evaluation code from logged experimental artifacts and were not generated or altered by the AI tool. The AI tool was not used to design the CRL-FM architecture, run experiments, or generate any of the underlying numerical results. The authors have reviewed all AI-assisted text, tables, and citations for accuracy against the source artifacts and original publications, and take full responsibility for the final content of this work.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their time.

REFERENCES

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximally interfered retrieval. *Advances in Neural Information Processing Systems*, 2019.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. In *arXiv preprint arXiv:1902.10486*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. *International Conference on Machine Learning*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models. *International Conference on Learning Representations*, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *Proceedings of EMNLP*, 2021.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of ACL-HLT*, 2011.

Arun Mallya and Svetlana Lazebnik. PackNet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. *International Conference on Learning Representations*, 2023.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. In *arXiv preprint arXiv:1606.04671*, 2016.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, 2018.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. CODA-Prompt: COntinual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of EMNLP*, 2013.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. LAMOL: LAnguage MOdeling for lifelong language learning. *International Conference on Learning Representations*, 2020.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. *Findings of EMNLP*, 2023.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. DualPrompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022b.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*, 2024.

Enneng Yang, Li Shen, Zhenyi Wang, Shiwei Liu, Guibing Guo, and Xingwei Wang. Mixture-of-experts for continual learning: A survey. *arXiv preprint arXiv:2408.16101*, 2024.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28, 2015.